SEGMENT SELECTION CONSIDERING LOCAL DEGRADATION OF NATURALNESS IN CONCATENATIVE SPEECH SYNTHESIS

Tomoki Toda[†‡]*, Hisashi Kawai*[†]*, Minoru Tsuzaki*[†]*, and Kiyohiro Shikano*[‡]

[†]ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, "Keihanna Science City" Kyoto, 619-0288 Japan
[‡]Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0101 Japan

ABSTRACT

In this paper, we investigate the effect of using a novel cost, RMS (Root Mean Square) cost, for segment selection for concatenative Text-to-Speech. The RMS cost is affected not only by the total degradation of naturalness but also by the local degradation of naturalness. From the results of experiments comparing this approach with segment selection based on a conventional average cost, it is found that (1) in the segment selection based on the RMS cost a larger number of concatenations causing slight local degradation are performed in order to avoid concatenations causing greater local degradation and (2) the effect of the RMS cost has little dependence on the size of the corpus. Moreover, we clarify that the naturalness of synthetic speech can be slightly improved by utilizing the RMS cost.

1. INTRODUCTION

We are constructing a concatenative Text-to-Speech (TTS) system based on a large-sized corpus that has high quality and high coverage on both phonetic environment and prosody. However, in order to realize a higher and more consistent quality of synthetic speech, it is necessary to design a cost function that is suitable for the perceptual characteristics [1][2].

In conventional segment selection, the optimum segment sequence is selected by minimizing the average cost calculated as the sum of local costs. The average cost shows the degradation of naturalness over the entire synthetic speech [3][4]. However, it might be assumed that local degradation of naturalness would have much effect on the naturalness of synthetic speech. In order to investigate this issue, we have evaluated various functions to integrate the local costs in selecting the optimum segment sequence [5]. From the results of perceptual experiments, it has been found that the RMS (Root Mean Square) cost, which is affected by both the average cost and a maximum cost showing the local degradation of naturalness, has the best correspondence to the perceptual scores. Therefore, it is possible to select a better segment sequence by utilizing the RMS cost in the segment selection.

In this paper, we compare segment selection based on the RMS cost with that based on the average cost in detail. In order to clarify how the local degradation of naturalness is alleviated by utilizing the RMS cost, selected segment sequences are analyzed from various points of view. We also show the relationship between the effectiveness of the RMS cost and the size of the corpus. Furthermore, the results of a preference test on the naturalness of synthetic speech clarify which of the two costs can select the best segment sequence.

The paper is organized as follows. In Section **2**, cost functions for segment selection are described. In Section **3**, the effectiveness of utilizing the RMS cost is described, and the results of the subjective evaluation are given in Section **4**. Finally, we summarize this paper in Section **5**.

2. COST FUNCTIONS FOR SEGMENT SELECTION

2.1. Local Cost

A local cost shows the degradation of naturalness caused by each candidate segment. In our Japanese concatenative TTS system under development, the local cost is calculated as the weighted sum of five sub-costs [5]. In this paper, in order to represent this cost more easily, we divide these sub-costs into two commonly used costs, i.e. a target cost C_i^t and a concatenation cost C_i^c [3][4].

The local cost $LC_i(u_i, t_i)$ at a phoneme u_i is given by

$$LC_i(u_i, t_i) = w_t \cdot C_i^t(u_i, ti) + w_c \cdot C_i^c(u_i, u_{i-1}), \quad (1)$$

$$w_t + w_c = 1, \quad (2)$$

where t_i denotes a target phoneme. w_t and w_c show the weights for the target cost and the concatenation cost, respectively. If u_{i-1} and u_i are connected in the corpus, the concatenation cost $C_i^c(u_i, u_{i-1})$ becomes 0.

In our segment selection, concatenations at some phoneme centers, i.e. preceding vowel centers of voiced phonemes and unvoiced fricative centers, are also allowed in order to alleviate audible discontinuity [6]. This algorithm allows the utilization of both phoneme units and diphone units and is similar to the AT&T NextGen TTS system [7], which performs segment selection based on half-phoneme units. When concatenation is allowed at the phoneme centers, the local cost $LC_i(u_i^l, u_i^f, t_i)$ at the half-phoneme segments (u_i^f and u_i^l) for a target t_i that is divided into half-phonemes (t_i^f and t_i^l) is calculated as

$$LC_i(u_i^l, u_i^f, t_i) = w_t \cdot (w_i^f \cdot C_i^t(u_i^f, t_i^f) + w_i^l \cdot C_i^t(u_i^l, t_i^l)) + w_c \cdot (C_i^c(u_i^f, u_{i-1}) + C_i^c(u_i^l, u_i^f)), \quad (3)$$

$$w_i^f = \frac{dur(u_i^f)}{dur(u_i^f) + dur(u_i^l)},$$

$$w_i^l = \frac{dur(u_i^l)}{dur(u_i^f) + dur(u_i^l)}, \quad (4)$$

where $dur(u_i^f)$ and $dur(u_i^l)$ show duration of the first half phoneme u_i^f and that of the last half phoneme u_i^l, respectively. If a diphone unit is used, the concatenation cost $C_i^c(u_i^f, u_{i-1})$ becomes 0. Furthermore, if a phoneme unit is used, the concatenation cost $C_i^c(u_i^l, u_i^f)$ becomes 0.

Our segment selection does not allow concatenation between C and V (C: Consonant, V: Vowel) and utilization of the half phoneme segments. Therefore, minimum units preserve either the important transitions between phonemes or the characteristics of Japanese syllables comprised of CV or V.

2.2. Average Cost

As a function to integrate the local costs, the average cost AC is often used and is given by

$$AC = \frac{1}{N} \cdot \sum_{i=1}^{N} LC_i(u_i, t_i), \quad (5)$$

where N denotes the number of phonemes in the utterance. Minimizing the average cost is equivalent to minimizing the sum of the local costs for the selection.

2.3. RMS Cost

In order to select the optimum segment sequence by taking account into not only the total degradation but also the local degradation, we minimize the RMS cost, $RMSC$, which is given by

$$RMSC = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} \{LC_i(u_i, t_i)\}^2}. \quad (6)$$

Actually, only the sum of the square local costs is calculated for the selection.

3. EFFECTIVENESS OF RMS COST

We utilized 1131 utterances as an evaluation set in order to compare the segment selection based on the RMS cost with that based on the average cost. These utterances were not included in the corpus used for segment selection.

3.1. Effect on sub-costs

We found that in the segment selection based on the RMS cost, the standard deviation of the local cost is smaller than that of the segment selection based on the average cost, although the average of this cost is slightly worse [5]. In this section, we investigated the effects of both the target cost and the concatenation cost in order to clarify the effectiveness of adopting the RMS cost.

The target cost is shown in **Fig. 1** as a function of corpus size. The average of the target costs is degraded, and the standard deviation increases slightly by utilizing the RMS cost. **Figure 2** shows the concatenation cost as a function of corpus size. Although the averages of concatenation costs are equal between the average cost and the RMS cost, the standard deviation becomes smaller by utilizing the RMS cost. These results show that the effectiveness of decreasing the standard deviation of the local cost is dependent on the concatenation cost. On the other hand, the average of the local cost is slightly worse as a consequence of degradation of the target cost.

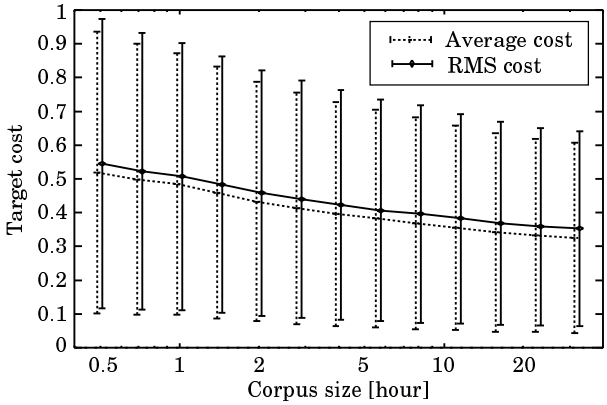

Fig. 1. Target cost as a function of corpus size.

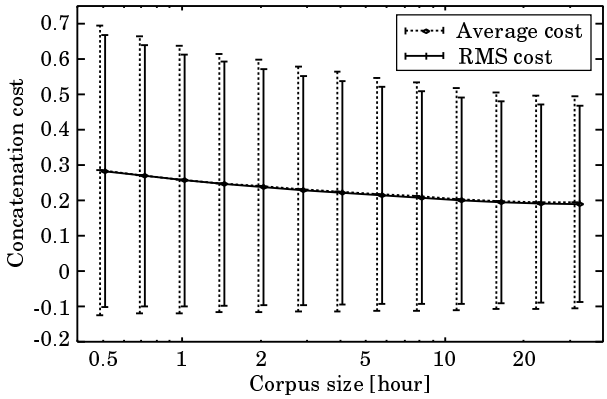

Fig. 2. Concatenation cost as a function of corpus size.

However, it is possible for these results to be influenced by the weights for sub-costs rather than by the local degradation, since we utilized the weight set in which the weight for the target cost is smaller than that for the concatenation cost. Therefore, we tried analyzing the effects of utilizing other weight sets. The same results were obtained for all weight sets, in which the ratios of the target cost to the concatenation cost were set to 1 to 2, 1 to 1.5, 1 to 1, 1.5 to 1, and 2 to 1. Therefore, the effectiveness as mentioned above depends not on the weights for sub-costs but on the function to integrate the local costs.

3.2. Effect on The Number of Concatenations

In order to clarify how the standard deviation of the concatenation cost is decreased, we investigated an effect on the number of concatenations.

The rate of increase in the number of concatenations is shown in **Fig. 3** as a function of corpus size. This rate is calculated by dividing the number of concatenations in the case of utilizing the RMS cost by those in the case of utilizing the average cost. By utilizing the RMS cost, the concatenation at boundaries between any phoneme and voiced phoneme decreases. However, the concatenations at boundaries between any phoneme and unvoiced phoneme and that at phoneme center increase. **Figure 4** shows the concatenation cost in each type of concatenation when the corpus size is

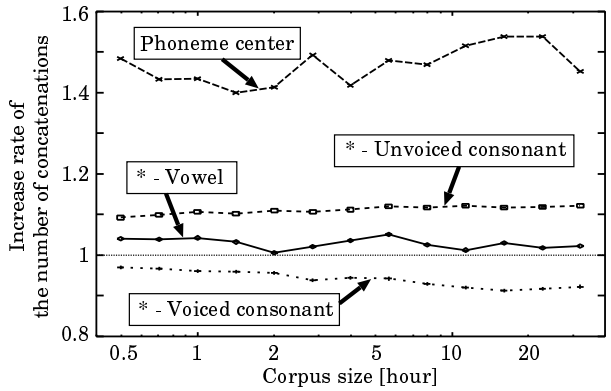

Fig. 3. Increase rate of the number of concatenations as a function of corpus size. "*" denotes any phoneme.

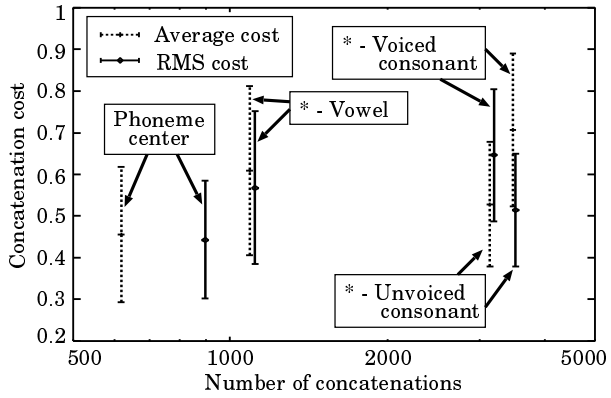

Fig. 4. Concatenation cost in each type of concatenation. The corpus size is 32 hours.

32 hours. The concatenation between any phoneme and unvoiced phoneme can often reduce the concatenation cost compared with that between any phoneme and voiced phoneme, since the former concatenation has no discontinuity caused by concatenating F_0s at a segment boundary.

These results show a tendency to avoid performing concatenations that often cause much local degradation of naturalness by instead performing more concatenations that cause slight audible discontinuity. As a whole, the number of concatenations increases rather than decreases. Therefore, a larger number of segments with shorter lengths, which only cause slight local degradation, are selected by utilizing the RMS cost [5].

3.3. Relationship of Cost Differences And Corpus Size

We investigated the differences in average costs, RMS costs, and maximum costs between the segment sequences selected by utilizing the average cost and those by utilizing the RMS cost. Then, the relationship between these differences and the corpus size was clarified.

The results are shown in **Fig. 5**. The maximum cost is shown as the largest local cost in each selected segment sequence. The cost differences are calculated by subtracting the costs of the segment sequences selected by utilizing the RMS cost from those by utilizing the average cost. From the results, the RMS cost works well for alleviating the local degradation of naturalness, since the maximum cost becomes small, i.e. the maximum cost difference is positive. Moreover, the differences in costs have little dependence on the corpus size. Therefore, the effect of the RMS cost can be found in any sized corpus.

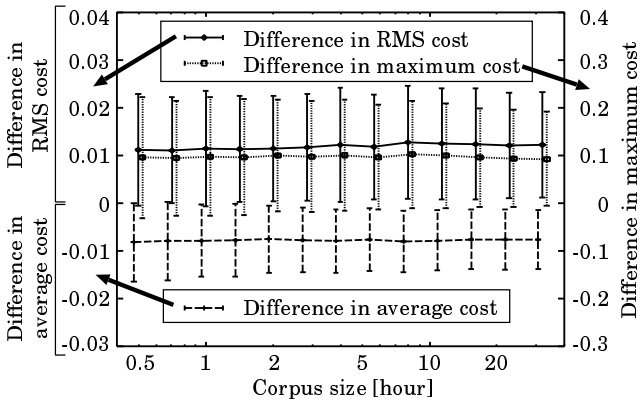

Fig. 5. Differences in costs as a function of corpus size.

4. SUBJECTIVE EVALUATION

4.1. Preference Test

We performed a preference test on the naturalness of synthetic speech to clarify which of the two costs could select the best segment sequence. The corpus size was 32 hours, and utterances used as test stimuli were not included in the corpus. Natural prosody was used as input information for segment selection.

The naturalness of synthetic speech was expected to be nearly equal between segment sequences having similar costs. Therefore, we used pairs of segment sequences that had greater cost differences in the test. Then, we selected the pairs with the larger differences in the average cost as well as those with the larger differences in the RMS cost. A scatter chart of the test stimuli is shown in **Fig. 6**. "Sub-set A" includes stimuli pairs with larger differences in RMS cost. These stimuli pairs are included in a region in which the normalized frequency of the RMS cost difference is less than about 20%. On the other hand, "sub-set B" includes stimuli pairs with larger differences in average cost, and likewise these are included in a region in which the normalized frequency of the average cost difference is less than about 20%. There were 20 stimuli pairs in each sub-set, and the total number of stimuli pairs was 35, since 5 pairs were included in both sub-sets. Eight Japanese listeners participated in the experiment. In each trial, synthetic speech by the segment selection based on the average cost and that by the segment selection based on the RMS cost were presented in random order, and listeners were asked to choose either of the two types of synthetic speech as sounding more natural.

The results **Fig. 7** show that the segment selection based on the RMS cost can synthesize speech more naturally than that based on the average cost in all cases: utilizing all stimuli, stimuli in sub-set A only, and stimuli in sub-set B only. However, this improvement is only slight.

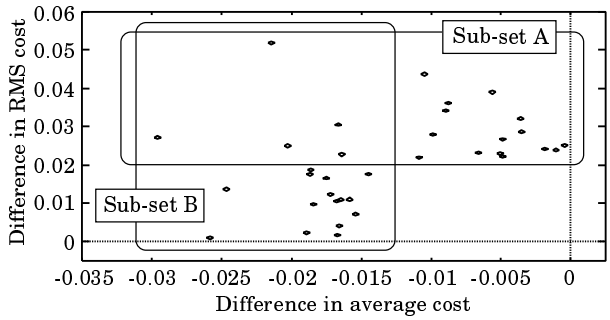

Fig. 6. Scatter chart of selected test stimuli. Each dot denotes a stimuli pair.

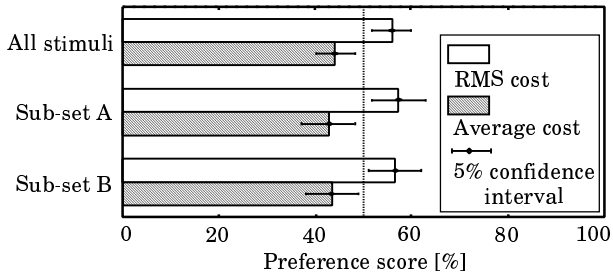

Fig. 7. Preference score.

4.2. Correspondence to Perceptual Score

We clarified correspondence of the RMS cost to the perceptual scores when the size of the corpus was varied [5]. However, our TTS system utilizes a large-sized corpus in order to synthesize speech more naturally and consistently, so it is worthwhile to investigate this correspondence in a range of lower RMS costs.

We performed an opinion test on the naturalness of the synthetic speech in order to clarify this issue. Test stimuli were included in the region covered in the case of utilizing a 32-hour corpus, in which the RMS costs were less than 0.4. They were selected from a large number of utterances synthesized by varying the corpus size. This selection was performed under the restriction that the number of phonemes in an utterance, the duration of an utterance, and the number of concatenations were roughly equal among the selected stimuli. The number of selected stimuli was 160. Eight Japanese listeners participated in the experiment. They evaluated the naturalness on a scale of seven levels. The perceptual score was calculated as an average of the normalized score calculated as a Z-score (mean = 0, variance = 1) for each listener in order to equalize the score range among listeners.

The correspondence of the RMS cost to the perceptual scores of the RMS cost is shown in **Fig. 8**. The correspondence is much worse than that in the case of utilizing stimuli that cover a wide range of the cost (correlation coefficient $= -0.843$) [5].

5. CONCLUSIONS

In this paper, we proposed segment selection that considers not only the degradation of naturalness over the entire synthetic speech but also local degradation. In this selection, the optimum segment sequences are selected by minimizing the RMS (Root Mean Square) cost instead of the average cost. From the results of experiments comparing this approach with segment selection based on the average cost, it was found that in segment selection based on the RMS cost a larger number of concatenations causing slight local degradation were performed in order to avoid concatenations causing greater local degradation. Moreover, the effectiveness of this selection was found for any sized corpus. We also performed subjective experiments on the naturalness of synthetic speech. As a result, it was clarified that the naturalness of synthetic speech can be slightly improved by utilizing the RMS cost. However, the correspondence of the cost to perceptual scores is not adequate. Therefore, we need to design a cost function that can capture perceptual characteristics more accurately in order to improve the quality of segment selection and make it more consistent.

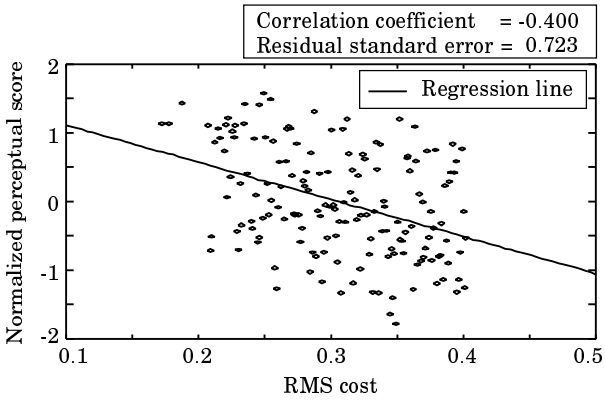

Fig. 8. Correspondence of RMS cost to perceptual scores.

Acknowledgment: This research was supported in part by the Telecommunications Advancement Organization of Japan.

6. REFERENCES

[1] H. Peng, Y. Zhao, and M. Chu, "Perpetually optimizing the cost function for unit selection in a TTS system with one single run of MOS evaluation," Proc. ICSLP, pp. 2613–2616, Denver, U.S.A., Sept. 2002.

[2] M. Chu and H. Peng, "An objective measure for estimating MOS of synthesized speech," Proc. EUROSPEECH, pp. 2087–2090, Aalborg, Denmark, Sept. 2001.

[3] A. Black and N. Campbell, "Optimizing selection of units from speech database for concatenative synthesis," Proc. EUROSPEECH, pp. 581–584, Madrid, Spain, Sept. 1995.

[4] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," Proc. ICASSP, pp. 373–376, Atlanta, U.S.A., May 1996.

[5] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano, "Perceptual evaluation of cost for segment selection in concatenative speech synthesis," IEEE Workshop on Speech Synthesis, Santa Monica, U.S.A., Sept. 2002.

[6] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano, "Unit selection algorithm for Japanese speech synthesis based on both phoneme unit and diphone unit," Proc. ICASSP, pp. 465–468, Orlando. U.S.A., May 2002.

[7] A. Conkie, "Robust unit selection system for speech synthesis," Joint Meeting of ASA, EAA, and DAGA, Berlin, Germany, Mar. 1999.
http://www.research.att.com/projects/tts/pubs.html